Meeting Report

SNPs, Haplotypes, and Cancer: Applications in Molecular Epidemiology

Timothy R. Rebbeck,[1] Christine B. Ambrosone,[2] Douglas A. Bell,[3] Stephen J. Chanock,[4] Richard B. Hayes,[4] Fred F. Kadlubar,[5] and Duncan C. Thomas[6]

[1]School of Medicine and Abramson Cancer Center, University of Pennsylvania, Philadelphia, Pennsylvania; [2]Roswell Park Cancer Institute, Buffalo, New York; [3]Environmental Genomics Section, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina; [4]National Cancer Institute, Bethesda, Maryland; [5]National Center for Toxicological Research, Jefferson, Arizona; and [6]University of Southern California, Los Angeles, California

Motivation

The ongoing discovery of single nucleotide polymorphisms (SNPs) and characterization of haplotypes in human populations is having a fundamental impact on molecular epidemiology. While likely common polymorphic variants interact with exposures to cause human cancer, the ability to evaluate the role of SNPs in human disease is limited by available methodologies. Numerous association studies of SNPs or haplotypes have been published to elucidate the etiology of cancer, but there has been inconsistency in the ability to replicate results. Therefore, a major goal of the field is to develop the analytical tools needed to examine the explosion of genetic information available for relating genetic variants to well-defined epidemiological end points. Inconsistent methodology and results from association studies have deflected attention away from the need to establish sound methodologies for both execution and interpretation of association study data. Therefore, we must optimize epidemiological, statistical, and laboratory approaches to achieve credible outcomes in association studies.

The goal of the AACR-sponsored conference ''SNPs, Haplotypes, and Cancer: Applications in Molecular Epidemiology'' was to address methodological developments for epidemiological studies investigating complex interrelationships of SNPs, haplotypes, and environmental factors with cancer. As summarized below, the content of this meeting included discussions related to the following:

1. *Gene Choice:* What genes and variants should be studied and how? What are the relative merits and costs of candidate gene *versus* genome-wide association studies? How should information about SNP function be incorporated into association studies?
2. *Laboratory and Genotype Data:* What are appropriate laboratory methods? How can adequate quality control be achieved? What is the role of public database information?
3. *Study Design and Analysis:* What study design and analysis approaches are required to achieve reproducibility among studies? What issues need to be faced when dealing with population genetics structure, including haplotype blocks and ethnic variability?

Choosing Genes and Haplotypes for Association Studies

Approximately ten million SNPs exist in the human genome, with an estimated two common missense variants per gene (*e.g.*, Ref. 1). At least 5 million SNPs have already been reported in public databases (2). However, the ability to apply these SNPs in association studies is limited by problems validating a SNP's identity, characterizing its occurrence in relevant populations, and understanding its function. Likely, only a small subset (perhaps 50,000–250,000) of the total number of SNPs in the human genome will actually confer small to moderate effects on phenotypes that are causally related to disease risk (3).

At the meeting, Stephen Sherry (National Center for Biotechnology Information) suggested several classes of variants to consider, including SNPs, deletion/insertion polymorphisms (DIP), simple tandem repeat (STR) polymorphisms, named polymorphisms (*e.g.*, Alu/− dimorphisms), and multinucleotide polymorphisms (MNP). Of these, ~3 million SNPs are estimated to be within or 2 kb upstream or downstream from a gene. Sherry reported a ''snapshot'' of gene-centric SNPs in the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP) as of September 2003. The distribution of these variants was 63% intronic, 11% untranslated region, 1% nonsynonymous, 1% synonymous, 24% locus region, <1% splice site, and <1% unknown coding variant. Recent surveys of human genetic diversity have estimated that there are about 100,000–300,000 SNPs in protein coding sequences (cSNPs) of the entire human genome (1). cSNPs are of particular interest because some of them, termed nonsynonymous SNPs (nsSNPs) or missense variants,

Cancer Epidemiol Biomarkers Prev 2004;13(5):681–7

Received 12/8/03; revised 1/14/04; accepted 1/22/04.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Report of a conference held at Key Biscayne, FL, September 13–17, 2003.

Requests for reprints: Timothy R. Rebbeck, Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, 904 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021. Phone: (215) 898-1793; Fax: (215) 573-2265. E-mail: trebbeck@cceb.med.upenn.edu

introduce amino acid changes into their encoded proteins. nsSNPs constitute about 1% of all SNPs. The rarity of nsSNPs may be a consequence of selective pressures. However, a significant fraction of functionally important molecular diversity in the human population likely is attributable to the effects on protein function caused by nsSNPs or alterations in the regulation of genes (known as rSNPs). For example, the kinetic parameters of enzymes, the DNA binding properties of proteins that regulate transcription, the signal transduction activities of transmembrane receptors, and the architectural roles of structural proteins are all susceptible to perturbation by nsSNPs and their associated amino acid polymorphisms. Similarly, John Potter (Fred Hutchinson Cancer Research Center) argued that perturbations in the regulation of key elements of a pathway could influence the risk for cancer outcomes either directly or through interacting pathways. Tom Hudson (McGill University) added that a major challenge of future studies will be to identify and characterize regulatory variants. The researcher will have to ask whether the regulatory SNP itself or the haplotype in which it may be imbedded is really the functional unit of interest.

The identification of biologically meaningful, disease-causing variants from among this large amount of genomic variability is a key challenge for association studies. Joel Hirschhorn (Massachusetts Institute of Technology) invoked the common disease, common gene hypothesis (4, 5) using methods that rely on knowledge of candidate genes or methods that rely on linkage disequilibrium (LD). Other publications have espoused the importance of rare variants as determinants for common diseases (4). Candidate gene approaches have the advantage of maximizing inferences about biological plausibility and disease causality. However, candidate approaches are limited by the amount of information that is available about the function of the gene in a specific disease process. Alternatively, genome-wide approaches have the advantage of scanning the entire genome for associations without having to rely on choosing *a priori* candidates. With advances in high-throughput technology and genome-wide association methods, these approaches will be more tractable than in the past. Stephen Chanock (National Cancer Institute) and David Hunter (Harvard University) stated that the candidate gene approach remains viable despite some limitations. The expectation of genome-wide approaches is still several years away because of formidable issues of cost and availability of genotyping platforms and analytical programs. Hunter, in a talk entitled ''Death, Taxes, and Candidate Genes,'' further stated that regardless of the initial approach, research must ultimately result in candidate gene studies to identify biologically meaningful causal associations involving specific genes.

Candidate Gene and SNP Choice. A major challenge in candidate gene studies is to choose appropriate candidate genes usually based on sound and plausible biologically driven hypotheses. Chanock and Stacey Gabriel (Massachusetts Institute of Technology) stressed choosing markers based on (*a*) strong prior information about biological pathways or linkage data; (*b*) functional correlates for a SNP or haplotype, including pathway or the use of evolution-based approaches to identify related genes based on sequence homology or gene family; and (*c*) SNP haplotype studies that start with a ''simple'' haplotypes (often including known nsSNPs or rSNPs), which can be expanded to increase the density of SNPs across the haplotype. Regardless of the approach for choosing markers, validation of associations in both comparable and different genetic backgrounds will be required. At the same time, the working hypothesis will likely become increasingly complex as knowledge of interrelated pathways is considered to account for relevant biological interactions.

William Evans (St. Jude's Children's Hospital), Gareth Morgan (University of Leeds), and Richard Weinshilboum (Mayo Clinic) presented the pharmacogenetics and pharmacogenomics paradigm for studies of candidate gene and SNP identification, gene discovery, and genotype-environment interaction. Pharmacogenetics and pharmacogenomics are excellent paradigms for studies that extend beyond etiology to studies of treatment response, gene expression changes, survival, side effects or toxicities relating to specific agents, timing of later events, and dosing. With respect to candidate genes, Weinshilboum stated that the paradigm for functional gene discovery in pharmacogenetics began by using the distribution of phenotypic traits to infer genetic effects. More recently, it has been possible to relate functionally significant DNA sequence variation to clinically important variability. Both of these approaches are complementary and should be both done to understand the functional significance of genes and SNPs. Evans also stated that gene expression profiling can be valuable to identify and characterize candidate genes (*e.g.*, for treatment response). Evans also presented examples in which genetic profiles differed by exposure (*i.e.*, where combinations of drug treatments did not evoke the same expression profile as each treatment individually). Therefore, expression profile approaches may be useful for identification of novel genes, characterizing function, novel disease classifications, and studying genotype-environment interactions.

Genome-Wide and Haplotype-Based Association Approaches. Gabriel and Eric Lai (GlaxoSmithKline) noted that causal (candidate) variants need not be studied directly but that gene discovery studies can be accomplished using a strategy that relies on LD between genetic variants. This represents the underlying premise behind whole genome SNP scans. The whole genome association approach can identify new candidate genes or regions. Millions of SNPs are available from the ATLAS project (http://www.confirmant.com/indexns4.html), the SNP Consortium (http://snp.cshl.org), and through public databases such as dbSNP (http://www.ncbi.nlm.nih.gov/SNP). Over >150 million SNPs are expected to be analyzed. These data can be used to define haplotype block structures across the genome and thus facilitate selection of SNPs for whole genome analysis.

Lai and Gabriel outlined approaches for undertaking genome-wide or haplotype-based studies. First, appropriate epidemiological study designs and adequate statistical power are essential. The number of samples and the number of SNPs required for these studies may be substantially higher than typical studies of the past. Second, results of genome-wide SNP association studies might not be easily replicated in subsequent studies but

could still identify causative regions of the genome. Similarly, large-scale genome-wide scans may find surrogate markers that will distinguish cases from controls but may not identify causative SNPs. Therefore, replication of associations is crucial to lead to valid and causative associations. Third, LD blocks exist throughout the genome, but these blocks are of varying length and appear to vary according to differences in population genetics. For example, Stephen O'Brien (NCI) stated that LD block size tends to be shorter in individuals of African ancestry and longer in Caucasians. O'Brien also reported that ~400,000 conserved sequence blocks exist in the human genome, which have been established by evolutionary constraints, specifically across species. Within these LD blocks, there is strong allelic association and limited haplotype diversity. Where haplotype diversity exists, particularly informative SNPs that best characterize a haplotype (tagSNPs) can be used to limit the amount of laboratory and analytical work in haplotype-based studies. Fourth, use of haplotype block information has been proposed to increase power 15–50% compared with a SNP-based analysis (6, 8). However, complete (and resource-intensive) studies of SNPs in a region are required to achieve sufficient statistical power. The alternative of studying incomplete sets of SNPs in a genomic region may result in less power but still identify causative loci. In this regard, several questions remain: What level of genomic coverage (*i.e.*, how many SNPs) is required to achieve an adequate result? Are tagSNP approaches adequate? How well do haplotype blocks need to be characterized and in what populations before tagSNPs can be reliably used? As an intermediate approach, Gabriel suggested a survey approach of candidate genes that encompassed 100 kb surrounding 200 candidate genes with 1 SNP (each with >5% minor allele frequency)/5 kb.

After introductory remarks justifying a common SNP haplotype-based association approach toward assessing risks associated with genomic variation in candidate genes, Daniel Stram (University of Southern California, Los Angeles, CA) introduced a formal statistical measure (R_h^2) of the predictability of haplotypes based on genotypes and described the use of this criterion for optimally picking haplotype tagging SNPs to be genotyped in large case-control studies. Stram described two approaches for estimating haplotype-specific risks in case-control studies (7, 8). The first is to compare haplotype frequencies in cases and controls separately. Analysis must allow for error in estimates of haplotype frequencies and generally cannot control for covariates. Second, regression substitution methods exist (9), in which an expected haplotype ''score'' is calculated as a predictor variable (*e.g.*, using proc haplotype in SAS) as if it were equivalent to the true haplotype. Under the alternative hypothesis, bias in estimates of effects can be evaluated and the degree of bias related to the formal measure of haplotype uncertainty (*i.e.*, R_h^2). Stram presented data that demonstrated biases in using the regression substitution methods are small to none if an adequate set of haplotype tagging SNPs is studied. However, biases increase as fewer haplotype tagging SNPs are included in analysis. The development of these and other computational approaches to study haplotype data in samples of unrelated individuals (*e.g.*, in case-control or cohort studies) will facilitate the evaluation of haplotypes in association studies.

Conclusion: Genome-wide association approaches should be feasible in the years to come but provide formidable challenges in throughput, databasing, and analysis. Candidate gene approaches will remain critical to confirm causal relationships of specific genes in regions identified by genome-wide or LD-based approaches. Moreover, the lessons learned from genome-wide studies are readily applicable to candidate gene studies and vice versa. Similarly, while knowledge of the functional significance of SNPs is key to understanding the biological basis of an epidemiological association, function can be determined in advance for candidate gene studies or after the identification of novel genes from genome-wide association studies.

Laboratory Approaches and Genome Database Resources

Optimizing Laboratory Throughput and Quality Control. Doug Bell (National Institute of Environmental Health Sciences) noted that high parallel (*i.e.*, many genotypes, few samples) and high throughput (*i.e.*, few genotypes, many samples) approaches are becoming widely available to molecular epidemiological studies. Several speakers proposed that an optimal genotyping approach should include 5–10% duplicate samples, 5% of SNPs to be genotyped on both DNA strands, <5% no call rate, >99% accuracy (*e.g.*, using validity checks with family sets), and be designed to require the smallest number of primers possible. Lai and Gabriel indicated that low-cost approaches (*e.g.*, per genotype cost as low as $0.04) are theoretically possible, but this estimate is not realistic for most laboratories. The largest single cost is that of hardware, although primer cost mounts with increasing the number of assays to be performed. The appropriate way to calculate genotyping cost is to determine total money spent (including hardware, reagents, and labor) divided by the number of genotypes generated.

A major issue for laboratory approaches in association studies is quality control. Hirschhorn, O'Brien, Gabriel, and others noted that genotype misclassification can result in bias (*e.g.*, toward the null hypothesis for nondifferential misclassification). Common laboratory problems include DNA contamination, inadequate quality or quantity of DNA, ''misarray'' of samples/plates, and assay error. Several quality control measures were suggested to address these issues. Chanock and Gabriel suggested that sample handling error may be reduced by typing microsatellite markers (fingerprinting) or otherwise genetically ''barcoding'' study participants' samples to detect contamination and identify sample mix-ups. For example, the Identifiler system (Applied Biosystems, Inc., Foster City, CA) consists of 15 microsatellite markers, which can be typed on every sample to monitor contamination or mishandling. Similarly, addition of parent-offspring trios to plates of DNA can help to identify non-Mendelian transmissions, which may reflect sample mix-ups or contamination. Estimation of Hardy-Weinberg proportions and tests of Hardy-Weinberg equilibrium can identify deviations from expected

proportions. Blind duplicates and blanks should be included in every genotyping run. For example, the ''360 rule'' dictates that 24 controls be included in every 384-well plate. Cases and controls should never be separated during the genotyping process to minimize the potential for differential genotyping error. Consideration of appropriate DNA extraction methods should be given to ensure appropriate DNA quality and quantity. Finally, a critical additional approach for ensuring high-quality laboratory data is the appropriate use of bioinformatics approaches to data handling to minimize error in sample handling genotype data. The use of a Laboratory Information Management Systems (LIMS) is critical for reducing data errors, particularly when these systems include built-in error checks.

DNA pooling can be used to increase the efficiency of association studies. Pak Sham (King's College) showed that for N subjects and M SNPs, $N \times M$ genotypes are required without pooling but $2 \times M$ genotypes are required for cases and controls using pooling (not considering replicate pools, which may also be required to account for pool construction error). Accurate sample dilution and quantitation are required to obtain equal DNA amounts from each person in pool. Pool construction may therefore be time consuming and expensive because extensive DNA quantitation is required. For example, Sham estimated that it requires one technician week to aliquot, dilute, and quantitate 100 samples. Pool construction error can occur if aliquoting the identical amount of DNA from each sample is not achieved. Pool measurement error can occur if inaccurate measuring the allele frequency is made from the pools. A potential quality control approach is to make multiple duplicate pools. Sham reported that pool construction error is much smaller than pool measurement error (although this may vary from laboratory to laboratory). As measurement error increases, information in the pool decreases and optimal pool fraction decreases. Therefore, pool size must be balanced against the number of measurements to be made: larger pools may not save as much effort because many more measurements may be required to minimize error. Lai noted that while DNA pooling may be a useful tool for screening associations, pooling strategies can be limited because once pools are constructed, reassignment of phenotype or removal from the pool is impossible. In addition, pools can only be made once all of the samples are all in hand. This requires a completed study before pooling approaches can be undertaken. Despite these potential limitations, DNA pooling serves as an example of the efficient approaches that may be required to undertake large-scale association studies.

Public Databases and Resources. A large amount of genomic information is available on public databases that can be of value to researchers undertaking association studies. For example, Sherry presented the information available via dbSNP (http://www.ncbi.nlm.nih.gov/SNP). Meredith Yeager (NCI) and Sherry reported that most genome information comes from data mining and genome assemblies; thus, the potential for database errors is large. For example, SNPs reported in 5–16% of coding regions represent paralogous variants that are due to duplicated segments (duplicons) and are therefore not real SNPs (10). Similarly, 15–30% of SNPs are not verified and may not exist. In general, SNP frequency estimates are not widely available, and when they are, they are often based on little information (few individuals). Because SNP frequencies can vary substantially by ethnicity, SNP frequencies may not be useful if frequencies are not reported by ethnicity. Only a few public SNP databases report sequence-verified SNPs. In addition, many SNP assays are not validated. Those assays that are validated sometimes provide inconsistent results by various genotyping methods, including sequencing. Many layers of validation, including laboratory methods and ethnicity-specific frequency estimation in multiple ethnic groups, are required to properly annotate a SNP before assays can be reliably designed. Tim Hubbard (Sanger Centre) also reported that databases such as ENSEMBL (http://www.ensembl.org) can be used to determine what genes have been validated and includes functional annotation. The ENSEMBL database also allows cross-species comparisons and data mining tools to explore evolutionary conservation and an interface for data analysis by individual investigators using their own data and/or public data.

To illustrate the potential for error in relying solely on public databases, Yeager studied 480 candidate genes. She and her colleagues observed 7.8 SNPs/gene that passed the annotation and assay validation protocol. This translates to ~1 probable SNP/154 bp and represents a higher density than previously reported. Thirty-nine percent of SNPs identified by resequence analysis of 204 chromosomes were not previously reported in dbSNP. These data suggest that the density of unknown SNPs is higher than expected, and many are not reported in public databases. If this is the case, designing assays will be prone to failure when unknown SNPs lie within the PCR product of interest or are within the probe being used and are not considered in assay design. This may lead to inability to design assays or result in inaccurate assays. This kind of variability in assay design and execution may induce genotyping error that contributes to inconsistency of results among association studies.

Conclusion: Quality control measures must be implemented to minimize genotype error that may result in bias and irreproducibility in association studies. Standardized approaches for biosample handling, genotyping, data collection, and data processing quality control methods should be developed. This can include systematic opportunities for laboratory exchange to cross-validate or replicate genotyping and assess error rates.

Study Design and Statistical Analysis: In Search of a Believable Result

Are Association Studies Replicable? A key issue in association studies is the ability to replicate association study findings. Replication of association studies is required not only to identify biologically plausible causative associations but also to conclude that a candidate gene has no meaningful etiological effect. Hirschhorn observed that most associations are not replicated. This lack of replication can be explained by false-positive reports (*e.g.*, spurious associations), false-negative reports (*e.g.*, studies that are insufficiently powerful to identify the association), or actual population differences

(*e.g.*, the true associations are different because of differences in genetic background, exposures, etc.). Given the perceived lack of consistency in association studies, what level of confidence can we have in associations reported to date?

To address this question, Hirschhorn reported on his group's meta-analysis (11) that included 25 inconsistent associations and 301 ''replication'' studies (*i.e.*, by ignoring the initial positive report). Most initial associations were not replicated, but an excess (20%) of replicated associations were seen when 5% were expected under the null hypothesis. This replication is not solely due to publication bias, because one would have to hypothesize that 40–80 negative studies were not reported rather than the average of 12 reported studies/association. Hirschhorn also concluded that it was unlikely that these replications represented false-positives due to ethnic stratification. Different LD patterns or other population patterns or population-specific modifiers (genes and/or environments) could also explain lack of replication, but this was unlikely to be a significant source of study inconsistency. The first positive reports also tended to be unreliable estimates for subsequently reported odds ratios (12), perhaps due to the ''Winner's Curse'' phenomenon, which predicts that the initial positive report overestimates the ''true'' value. Indeed, 23 of 25 associations studied showed evidence for a ''winner's curse.'' An additional consequence of this phenomenon is that replication studies may therefore require larger sample sizes because the actual replication effects may be smaller than suggested by the initial report. Despite these limitations, these data indicate that many associations are replicable and may therefore represent truly causative effects on disease.

Factors that Influence Association Study Results. To achieve believable, replicable association results, investigators must consider factors that influence the design, analysis, and interpretation of these studies. Standards must be established for agreeing on when associations do or do not exist based in part on the issues outlined below.

Etiological Complexity. Numerous speakers stated that the etiology of human disease is complex and the diseases themselves are etiologically heterogeneous. Therefore, association study methods need to address this complexity. Sholom Wacholder (NCI) and Hunter noted that considering interaction *versus* stratified effects may detect different kinds of effects and determine the context in which a gene's effect is likely to be important. The power and efficiency of association studies may be improved if genetic effects are studied in groups defined by exposure status, particularly in genotype-environment interaction studies. Other approaches that can help to address complexity include studying subsets of cases defined by histopathology or other characteristics to decrease heterogeneity, using methods that allow multidimensional classification of outcome, or by taking advantage of intermediate end points (*e.g.*, preneoplastic events and time to progression) rather than limiting studies to genotype-disease associations.

Power and Sampling Design. Studies with larger sample sizes have advantages over smaller studies, including pooled or meta-analyses of smaller studies. Potter stated that large cohort studies have advantages over individual case-control studies, but these must have adequate biosamples, risk factor data, and power to evaluate disease subtypes. In this context, he proposed to develop ''The Last Cohort,'' which would include 500,000 or more individuals with complete data and biosamples, to be followed longitudinally. Several paradigms for undertaking large studies include large prospective cohort studies, multicenter case-control studies (*e.g.*, InterLymph; Nat Rothman, NCI), consortia of individual cohorts or case-control studies (*e.g.*, the NCI Cohort Consortium for Breast and Prostate Cancer; Hunter), large cohort studies (*e.g.*, UK Biobank, http://www.ukbiobank.ac.uk; EPIC, http://www.iarc.fr/EPIC; Elio Riboli), and population-based family designs (*e.g.*, deCODE, http://www.decode.is; Laufey Amundadottir). Despite these existing data sets, genotyping technology and statistical methods have continued to develop, but the population resources required to address relevant research questions of interest have not advanced as quickly.

Despite the general desire to see large-scale studies, mounting them is challenging. Results from very large studies have the potential to generate small and clinically ''unimportant'' results that may not be reproducible unless a similar sample is studied. Riboli noted that while large studies have the advantage of being able to detect small magnitude effects and may minimize the chance of false-positive results, numerous smaller studies may be more efficient for gene discovery and replication than a few very large studies. Multicenter studies of relatively rare cancers may be required, particularly if subset analyses are to be done. However, multicenter studies raise additional concerns (especially if study consortium is developed post hoc), including consistency of genotyping and questionnaire data across centers, correlation and other confounding by center, and other study design and data collection differences across centers. The implications for meta-analyses of data are also unfavorable if there are publication biases or other methodological problems with the individual studies. Therefore, several issues need to be resolved before large-scale studies can be appropriately undertaken.

Population Structure. Inability to achieve replication among association studies may be due to characteristics of the study population. Hunter raised the concern that replication requires that studies themselves be comparable (*e.g.*, in terms of ethnicity or other confounding factors), which is not usually the case. Numerous speakers addressed the question of genetic structure of populations related to ethnicity or race. Jonathan Pritchard (University of Chicago) noted that Europeans are genetically the least diverse and thus unlikely to pose major problems in stratification and bias. Rick Kittles (Howard University) reported on population genetic structure that exists among admixed groups, particularly African Americans who demonstrate high genomic heterogeneity due to recent population admixture. The concern that arises out of the existence of this population structure is that confounding by ethnicity (*i.e.*, population stratification) may lead to improper inferences from association studies. Several individuals (Wacholder; Peter Shields, Georgetown University; Yiting Wang, University

of Pennsylvania) presented data suggesting that race and ethnicity are not important source of bias, particularly in North American populations of Western European descent when proper epidemiological methods are used because differences in baseline disease risks are not large enough to confer significant confounding. This view is consistent with a recent research by several investigators (13–16). For example, Cardon and Palmer argued that poor study design may be important than population stratification in conferring bias to association studies.

Several analytical approaches exist to either circumvent problems imposed by population genetic structure or that use this structure in gene identification. Pritchard suggested that population stratification can be addressed by using family-based methods, or by analyzing unlinked markers as controls. In the latter approach, if association is solely due to population structure, there should be similar associations across random markers across the genome. Several approaches exist that can assess this phenomenon (17, 18). The ''structured association'' approach identifies a set of individuals who are drawing their alleles from different background populations or ethnicities. This approach uses information from unlinked markers to infer their ancestry and learn about population structure. It further uses this population structure information to adjust the association that is observed. The ''genomic control'' approach instead uses the distribution of association tests statistics for the unlinked markers to adjust the usual χ^2 test of association for the overdispersion caused by hidden population stratification.

Taking advantage of population genetics structure can also aid association studies. O'Brien reported on the mapping by admixture LD approach, which reduces the number of SNPs that are needed for a whole genome association scan because LD tends to extend over a much broader range in recently admixed populations. This approach uses differences in disease risk by ethnicity and studies markers that also differ by ethnicity. Markers can then be studied to estimate the proportion of ethnic admixture (*e.g.*, proportion of African ancestry) using a hidden Markov model and then combine that analysis with mapping data to identify genomic regions that appear to differ between cases and controls. Similarly, Duncan Thomas (University of Southern California) discussed coalescent, Bayesian clustering, and haplotype sharing approaches to make inferences about the evolutionary history of SNP and haplotype distributions (8). These approaches can show how putative susceptibility alleles have arisen in specific populations and may thereby identify disease-associated haplotypes.

Data Interpretation. Assuming study design and analysis issues have been appropriately managed, it remains problematic to interpret results of many association studies to conclude that a SNP is (or is not) causally associated with disease. Several approaches have been suggested to objectively evaluate the evidence of a particular hypothesis. Wacholder posed this question by stating that three possible decisions can be made at the time an association is reported: the association is ''noteworthy''; no important association can exist; or no decision can be made. Wacholder proposed a false-positive report probability (FPRP) to aid in making this decision (19). The FPRP depends on prior probability, power, and size of effect. When the prior probability is high, the FPRP is low and an association is more likely to be correct. When prior probability is low, increasing sample size only marginally increases the chances of finding a true association. Thus, true associations are more likely to be identified if the prior probability of finding an effect is high. In addition, small studies may be more likely to give a false-positive result. The FPRP can be implemented to provide investigators with the ability determine the extent of toleration of false reports. Second, the investigator must choose a prior probability. For example, the investigator can ask what is the probability that there is a nontrivial causal effect (*e.g.*, relative risk = 1.5) between SNP and disease before performing data analysis. This prior probability can be chosen based on functional, genomic (*i.e.*, type of mutation and function), and epidemiological data (*i.e.*, incorporate previous reports and information about the quality of study, sample size, power, and relationship with diseases of likely similar etiology). Both Wacholder and Chanock indicated that future candidate gene studies may face lower prior probabilities as association studies expand to candidate genes about which less is known *a priori*. Third, the investigator must choose a clinically or etiologically meaningful effect size (*e.g.*, relative risk or odds ratio) and calculate the FPRP for each SNP and compute over a range of priors. The investigator can conclude that the result is ''noteworthy'' if the FPRP exceeds some predetermined threshold. Similarly, the investigator can compute a false-negative report probability to assist in deciding there is no association. As before, false-negative report probability depends on power, sample size, prior probability, and effect size.

Hirschhorn suggested the prior probability of association for a single random SNP depends on the existence of haplotype blocks, with prior probabilities in the range of 1 in 10,000 to 1 in 100,000. For candidate gene associations (assuming 300 candidates are studied, 3 haplotype blocks exist per gene, and there are 4 haplotypes/block, ~3600 candidate variants exist, and half of these are causal), the prior probability for association is 1 in 100 to 1 in 1000. Thus, candidate gene association studies are predicted to have a substantially higher prior probability of seeing an effect than random SNP association studies.

Conclusion: Appropriate study design and adequate statistical power are crucial to obtaining meaningful association results. Additional considerations of etiological complexity and heterogeneity, prior probability or FPRP of association, and population genetics structure should be incorporated in association studies. Central coordination of positive and null association reports would help researchers to digest the literature, but no such coordinated database currently exists to track this information.

How Can Meaningful Results Be Obtained from Association Studies?

Several critical recommendations were made to improve the chances of reporting replicable, believable associations (suggested by several participants including Hunter; O'Brien; Potter; Tom Sellers, Moffitt Cancer

Center; Daniela Seminara, NCI; and others). These recommendations should be incorporated into the design, analysis, and interpretation of all future association studies.

- Strive to decrease false-positive results. Consideration of FPRP may aid this process.
- Strive to decrease false-negative results. Perform studies using sufficiently large sample sizes to ensure adequate statistical power to detect small effect sizes and to study appropriate interactions or effects within relevant strata.
- Conclude associations exist only in the presence of small *P* values, possibly adjusted for multiple tests.
- Associations must be replicated to be believable. This may require pooled/meta-analyses, a coordinated interdisciplinary approach to speed up research pace that benefits from an economy of scale, maximized infrastructure for bioinformatics and biospecimen management, and increased data resource sharing.
- Identify criteria that can be used to decide when a SNP, haplotype, or gene is NOT associated with disease.
- Epidemiological associations should be coupled with biological function to better motivate studies and to enable interpretation of association results.
- Effects of population genetics structure should be further studied, including studies undertaken in well-characterized ethnic subsets when population structure may affect the results.
- Studies should consider knowledge of the prior probability that a SNP or haplotype is associated with the disease and increase this probability by selecting candidates from regions of linkage peaks, incorporating functional information, using bioinformatics computational tools.
- Use haplotype mapping approaches to identify genomic regions associated with disease followed by candidate gene studies to elucidate the actual causative genes and variants of interest. Similarly, extend candidate gene association studies using haplotype information.
- Laboratory approaches must optimize precision and accuracy, including appropriate quality control, sample preparation, and duplicate samples, and ensure sample integrity using DNA fingerprinting, robust genotyping assays, plating cases and controls together.
- Consider a full range of etiological models: for example, analyses should not be limited to only dominant or recessive effects unless strongly dictated by functional data. Gene dosage effects should also be considered.
- Consider the context of the association by evaluating environmental exposures (*i.e.*, genotype-environment interaction) and the epistatic context of associations (*i.e.*, genotype-genotype interactions). Methods should be developed for exploring the joint effects of many genes and environmental factors involved in a common pathway or competing pathways jointly rather than one at a time or in pairwise combinations (20, 5).
- Consider etiological heterogeneity by evaluating histopathological or other data that subclassify disease to get more homogeneous groupings for analysis.
- Novel designs and analytical methods should be considered, including multistage designs that combine information from haplotypes and population structure.
- Determine the clinical and population significance of a gene by assessing the relative and attributable risk seen in the association.
- Rules for data sharing, particularly when consortium studies are being undertaken, need to be established.
- Researchers should consider how to translate SNP-based association study results in terms of clinical and public health applications.

The ability to implement these and other measures will determine whether association studies involving SNPs or haplotypes will provide meaningful information about disease etiology or outcome and thus whether this information can be further translated into cancer prevention, clinical practice, or basic science studies that elucidate disease mechanism.

References

1. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999;22:231-8.
2. Salisbury BA, Pungliya M, Choi JY, Jiang R, Sun XJ, Stephens JC. SNP and haplotype variation in the human genome. Mutat Res 2003; 15;526(1-2):53-61.
3. Chanock S. Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. Dis Markers 2001;17(2):89-98.
4. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant. . .or not? Hum Mol Genet 2002; 11(20):2417-23.
5. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17(9):502-10.
6. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296(5576):2225-9.
7. Stram DO, Pearce CL, Bretsky P, et al. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered. In press 2003.
8. Thomas DC, Stram DO, Conti DV, Molitor J, Marjoram P. Bayesian spatial modeling of haplotype associations. Hum Hered 2003;56(1-3): 32-40.
9. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 2002;53(2):79-91.
10. Eichler EE, DeJong PJ. Biomedical applications and studies of molecular evolution: a proposal for a primate genomic library resource. Genome Res 2002;12(5):673-8.
11. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;3(2):177-82.
12. Ioannidis JP, Rosenberg PS, Goedert JJ, O'Brien TR. International Meta-Analysis of HIV Host Genetics. Commentary: meta-analysis of individual participants' data in genetic epidemiology. Am J Epidemiol 2002;156(3):204-10.
13. Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. Am J Hum Genet 2002;71(2):304-11.
14. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003;361:598-604.
15. Millikan RC. Re: Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2001;93(2):156-8.
16. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000;92(14):1151-8.
17. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55(4):997-1004.
18. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220-8.
19. Wacholder S, Chanock S, Garcia-Closas M, El ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96(1-3):434-42.
20. Conti DV, Cortessis V, Molitor J, et al. Bayesian modeling of complex metabolic pathways. Hum Hered 2003;56(1-3):83-93.